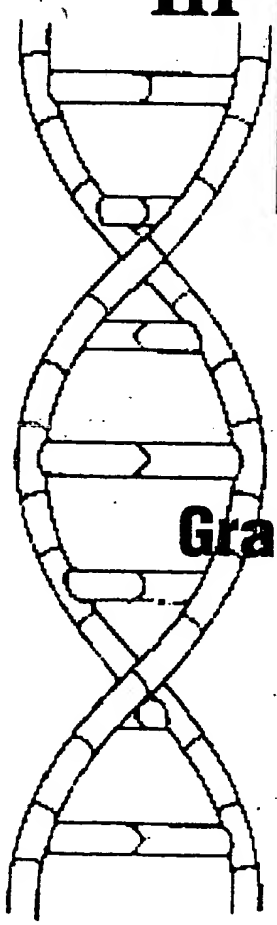
LABORATORY METHODS for the Detection of Mutations and Polymorphisms in DNA



Edited by
Graham R. Taylor

CRC Press
Boca Raton New York

Exhibit 1

Chapter 4

Microsatellite Analysis

S. J. Payne

4.1 Introduction

In the space of 5 years since the first report of their informativeness, thousands of microsatellites have been characterized and a high-resolution genetic linkage map of the human genome built. Microsatellites have been key tools in tracking disease genes both in clinical and research laboratories. Short tandem repeat loci (STRs) are used in forensics, identity testing, and in analysis of population structure (this will increase as the Human Diversity project expands). STR markers are abundant, highly polymorphic and technically very simple to analyze. It is not too big a cliché to say that microsatellites have revolutionized genetic analysis.

The first DNA polymorphisms to be exploited in genetic linkage studies were single base-pair sequence variations which occurred within restriction endonuclease recognition sites. These polymorphisms could be easily detected because variant sequences either created or abolished enzyme recognition sites and therefore resulted in restriction fragments of variable length. Botstein et al.[1] proposed constructing genetic maps using restriction fragment length polymorphisms (RFLPs). Although RFLPs are widely distributed throughout the genome, their utility is limited by low informativeness. Since most RFLPs are only dimorphic (either the enzyme cuts or it doesn't), the maximum heterozygosity of 50% can only occur when both alleles are equally represented in a population—most RFLPs have lower heterozygosities.

Concurrently with the development of RFLPs, another class of DNA polymorphism was characterized based on tandem arrays of repeated sequences.[2–4] Tandem sequence repetition is widespread in eukaryotic genomes and many types of repeat motif have been described. One common feature of repetitive sequence loci is that the number of repeat units differs between individuals, giving rise to arrays of variable length. Polymorphic markers based on variable numbers of tandem repeats (VNTRs) are potentially very informative because of the large number of alleles which may exist. The most polymorphic VNTRs ("minisatellites") have repeat units of between 12 and 60 or more base-pairs and a total array size of 0.5 to over 3 kb. The major limitation of minisatellite VNTRs is that they tend to be clustered at telomeres and are therefore of restricted value in constructing complete human genome maps.[4]

In the early 1980s a sub-class of repetitive loci were described with a repeat unit of only two base pairs—so-called "microsatellites".[5–8] It was not, however, until 1989 that the polymorphic nature of microsatellites was recognized.[9,10] As with larger VNTRs, microsatellites

0-8493-9231-4/97/$0.00+$.50

vary between individuals in the number of repeats in the array. Their nomenclature is informal and such loci are variously referred to as STRs, variable small sequence markers (VSSMs), simple sequence repeats (SSRs), dinucleotide repeats, CA blocks etc. The repeat unit may be from 1 to 6 bp and the most common microsatellite repeat motifs are A, AC, AAAN, AAN, AG, and AT[11] although the best characterized are dinucleotide (dC-dA/dG-dT) repeats. Microsatellites are extremely abundant, occurring with an estimated average frequency of one STR every 6 kb of human genomic sequence.[11] Microsatellites have clear advantages over the other polymorphisms described above. STRs often have multiple alleles and many have heterozygosity frequencies of 70% or more making them highly informative for genetic analysis. In addition, the loci are small enough to be analyzed using the polymerase chain reaction (PCR).[12,13] The significance of these factors was quickly recognized and microsatellites soon became markers of choice for many applications.

4.1.1 Informativeness of Microsatellites

The informativeness of a polymorphic marker depends upon the number of alleles and their relative population frequencies. In the context of genetic linkage studies (for example, predictive linkage analysis in a family with a genetic disease), the informativeness of a linked marker relates to the likelihood that the parental genotypes can be deduced following analysis of a child of an affected parent. Botstein et al.[1] described the polymorphism information content (PIC) which is a statistical assessment of informativeness of a marker. In order to evaluate a marker for PIC, firstly the frequencies of all possible genotypes for a given marker in a population and the frequencies of all mating-type combinations are estimated. Next, the probability of informativeness in offspring of each mating-type combination is calculated. Finally, a value for PIC is obtained by summing the mating-type frequencies multiplied by the probability of informative offspring.

Marker informativeness is more easily estimated by simply counting the number of heterozygotes in a suitably large sample set. PIC approximates to the observed frequency of heterozygosity. The greater the number of alleles at a given locus (and the more even the spread of allele frequencies in a population), the more informative will be the marker. This underlies the virtue of microsatellites in linkage analysis and gives measure to the extent to which microsatellites are much more informative than dimorphic systems such as RFLPs.

4.2 Applications

4.2.1 Construction of Genetic Maps

Genetic maps are constructed by linkage analysis. Linkage relationships (map order and distance between markers) are established by typing a collection of families with the markers of interest. Mapping information is obtained by detecting recombination between markers. Linkage analysis has been successful in mapping genes for a great number of inherited conditions as the first step in a positional cloning strategy. The disease itself is treated as polymorphic marker with alleles "mutant" and "normal." This clearly relies upon accurate clinical assessment of recombinant individuals in affected families.

4.2.2 Disease Gene Tracking

Clinical molecular genetics laboratories make use of linked marker to perform predictive or presymptomatic testing for at-risk individuals in affected families. Disease genes are tracked through families by analyzing inheritance of markers known to be closely-linked to the disease. Figure 4.1 shows an example of use of a STR to track an autosomal dominant, late

Exhibit 1